DOCUMENTOS DE TRABAJO

Modelling high frequency non-financial big time series with an application to jobless claims in Chile.

Antoni Espasa
Guillermo Carlomagno

N° 1023 Octubre 2024
BANCO CENTRAL DE CHILE

banco central Chile

Documentos de Trabajo del Banco Central de Chile
Working Papers of the Central Bank of Chile
Agustinas 1180, Santiago, Chile
Teléfono: (56-2) 3882475; Fax: (56-2) 38822311

Documento de Trabajo N° 1023

Working Paper N° 1023

Modelling high frequency non-financial big time series with an application to jobless claims in Chile.*

Antoni Espasa
Universidad Carlos III

Guillermo Carlomagno
Banco Central de Chile

Resumen

Este trabajo explora los desafíos de modelar series de tiempo de alta frecuencia, de grandes datos no financieros. Centrándose en datos diarios, horarios e incluso a nivel de minutos, el estudio investiga la presencia de diversas estacionalidades (diarias, semanales, mensuales y anuales) y cómo estos ciclos pueden interrelacionarse entre sí y ser influenciados por patrones climáticos y variaciones del calendario. Mediante el análisis de estas características cíclicas y las respuestas de los datos a factores externos, el trabajo explora el potencial de los modelos de cambio de régimen, dinámicos y no lineales para capturar estas complejidades. Además, propone el uso de Autometrics -un algoritmo automatizado para identificar modelos parsimoniosos- para dar cuenta conjuntamente de todas las peculiaridades de los datos. Los modelos resultantes, más allá del análisis estructural y la predicción, son útiles para construir indicadores líderes macroeconómicos cuantitativos en tiempo real, planificación de la demanda y estrategias de precios dinámicos en diversos sectores sensibles a los factores identificados en el análisis (por ejemplo, de servicios públicos, tiendas minoristas, tráfico o indicadores del mercado laboral). El trabajo incluye una aplicación a la serie diaria de solicitudes de cesantía en Chile.

Abstract

This paper explores the challenges of modelling high-frequency, non-financial big data time-series. Focusing on daily, hourly, and even minute-level data, the study in-vestigates the presence of various seasonalities (daily, weekly, monthly, and annual) and how these cycles might interrelate between them and be influenced by weather patterns and calendar variations. By analyzing these cyclical characteristics and data responses to external factors, the paper explores the potential for regime-switching, dynamic, and non-linear models to capture these complexities. Furthermore, it proposes the use of Autometrics –an automated algorithm for identifying parsimonious models– to jointly account for all the data's peculiarities. The resulting models, beyond structural anal-ysis and forecasting, are useful for constructing real-time quantitative macroeconomic leading indicators, demand planning and dynamic pricing strategies in various sectors that are sensitive to the factors identified in the analysis (e.g., of utilities, retail stores, traffic, or labor market indicators). The paper includes an application to the daily series of jobless claims in Chile.

∗ This paper was initially distributed as Tall big data time series of high frequency: stylized facts and econometric modelling. The views expressed in this paper are those of the authors and do not necessarily represent those of the Central Bank of Chile of its board members. The application in section 4 of the paper was developed within the scope of the research agenda conducted by the Central Bank of Chile (CBC) in economic and financial affairs of its competence. The CBC has access to anonymized information from various public and private entities, by virtue of collaboration agreements signed with these institutions. To secure the privacy of workers and firms, the CBC mandates that the development, extraction and publication of the results should not allow the identification, directly or indirectly, of natural or legal persons.

Modelling high frequency non-financial big time series with an application to jobless claims in Chile.*

Antoni Espasa[a] and Guillermo Carlomagno[b]

[a]Honorary Emeritus Professor, Universidad Carlos III, Madrid, Spain
(e-mail:antoni.espasa@uc3m.es)
[b]Central Bank of Chile (*e-mail:gcarlomagno@bcentral.cl*)

September 10, 2024

Abstract

This paper explores the challenges of modelling high-frequency, non-financial big data time-series. Focusing on daily, hourly, and even minute-level data, the study investigates the presence of various seasonalities (daily, weekly, monthly, and annual) and how these cycles might interrelate between them and be influenced by weather patterns and calendar variations. By analyzing these cyclical characteristics and data responses to external factors, the paper explores the potential for regime-switching, dynamic, and non-linear models to capture these complexities. Furthermore, it proposes the use of *Autometrics* –an automated algorithm for identifying parsimonious models– to jointly account for all the data's peculiarities. The resulting models, beyond structural analysis and forecasting, are useful for constructing real-time quantitative macroeconomic leading indicators, demand planning and dynamic pricing strategies in various sectors that are sensitive to the factors identified in the analysis (e.g., of utilities, retail stores, traffic, or labor market indicators). The paper includes an application to the daily series of jobless claims in Chile.

Keywords: *aggregation, several seasonality (daily, weekly, monthly and annual), complex annual calendar composition, weather variables, interactive effects, switching regimes, multiplicative, dynamic and non-linear structures, designing of exogenous variables, Autometrics, macroeconomic leading indicators, jobless claims.*

JEL: C01, C22, C55.

*This paper was initially distributed as Tall big data time series of high frequency: stylized facts and econometric modelling. The views expressed in this paper are those of the authors and do not necessarily represent those of the Central Bank of Chile of its board members. The application in section 4 of the paper was developed within the scope of the research agenda conducted by the Central Bank of Chile (CBC) in economic and financial affairs of its competence. The CBC has access to anonymized information from various public and private entities, by virtue of collaboration agreements signed with these institutions. To secure the privacy of workers and firms, the CBC mandates that the development, extraction and publication of the results should not allow the identification, directly or indirectly, of natural or legal persons.

1 Introduction

Big data has revolutionized our ability to collect and analyze information. It offers a vast amount of data points, often at much higher frequencies than traditional sources. This information, often referring to micro units and available at very high frequency, even in real-time, can be generated from web activity (Cavallo and Rigobon, 2016), sensor data (like traffic patterns), or smart meters. However, the very complexity of big data presents a challenge: transforming this information into usable knowledge requires significant processing (Franke and et al, 2016). This challenge has given rise to the field of Data Science, where Statistics and Computer Science play a crucial role.

This paper addresses the specific challenges posed by high-frequency, non-financial time series data. Adopting the term *tall big data* from Hendry (2015a), we characterize these datasets as having relatively few endogenous variables but potentially hundreds of exogenous variables—to be designed by the analyst—and a large number of observations. Collected at daily, hourly, or even minute frequencies, these data exhibit rapid fluctuations and intricate patterns. Unlike traditional monthly or quarterly data, they present distinct complexities related to seasonality, weather influences, and calendar effects.

Complex Seasonality. Traditional econometric models often account for basic seasonal patterns, such as monthly or yearly cycles. However, high-frequency data reveals a more nuanced picture. Daily, weekly, and even specific calendar day effects can become significant. Imagine, for example, analyzing hourly retail sales data. We might observe not just a weekly cycle (higher sales on weekends) but also daily fluctuations (higher sales during lunch hours) and potentially even variations based on the day of the week (higher sales on payday compared to Mondays). Capturing these intricate seasonal patterns is crucial for extracting meaningful insights from the data, as emphasized by Hendry (2015a).

The Influence of Weather and Calendar Effects. Beyond seasonality, high-frequency data can be significantly influenced by external factors like weather and calendar variations. For instance, daily electricity consumption might soar during heat waves, requiring adjustments in the model. Similarly, a holiday landing on a Monday could disrupt the typical weekly sales cycle. Understanding and accounting for these external influences is essential to avoid misleading conclusions from the data, as highlighted in the by Zhou et al. (2017).

Our modelling approach. This paper proposes a methodology for building econometric models that can handle the complexities of high-frequency, non-financial time series data. Our approach acknowledges the presence of multiple seasonal patterns, weather influences, calendar effects and outlying observations. We introduce a framework for constructing explanatory variables that capture these intricacies and adopt an estimation strategy that deals the potentially large number of explanatory variables. The resulting models, beyond structural analysis and forecasting, are useful for constructing real-time quantitative macroeconomic leading indicators, demand planning and dynamic pricing strategies in various sectors that are sensitive to the factors identified in the analysis (see e.g., Den Boer,

2015, Dutta and Mitra, 2017, Saharan et al., 2020). By capturing the complex interplay of factors influencing demand and customer behavior in real-time, these models could inform pricing decisions that adapt to fluctuations in seasonality, weather, and calendar effects. For instance, a utility company store might leverage such models to adjust prices based on real-time weather data, potentially offering lower rates during off-peak hours or periods of mild weather. Similarly, the widespread adoption of electronic price tags has enabled retail stores to implement dynamic pricing strategies, where weather and calendar effects can significantly influence pricing decisions.

Additionally, the proposed method provides an opportunity to "de-bug" the data itself. By identifying and adjusting for complex seasonality and weather effects, we can create a "cleaned" version of the high-frequency data. This cleaned data can then be used to construct real-time quantitative macroeconomic leading indicators, providing valuable insights for economic policymakers and businesses alike. Importantly, traditional signal extraction methods designed for lower-frequency (monthly/quarterly) data are ineffective even when applied to lower-frequency aggregates of high-frequency data. This is due to two primary reasons. Firstly, nonlinearities present at high frequencies cannot be accurately captured at lower frequencies, leading to biased estimates of the signals, as discussed in Section 3.2. Secondly, outliers are particularly prevalent in high-frequency data. When aggregated to a lower frequency, such as monthly, many outliers may be obscured or rendered statistically insignificant. However, these outliers can still influence the values of the lower frequency aggregated data, even if they do not appear as statistically significant anomalies. Thus, outlier correction should be performed at the high-frequency level.

While our research addresses similar challenges to Proietti and Pedregal (2023), particularly regarding complex seasonality and calendar effects, our approach offers two main advantages. **(i) Causal Foundations:** We leverage our understanding of the data's underlying characteristics to build an interpretable model. This ensures all factors included in the model have clear causal relationships, enhancing user comprehension. **(ii) Joint Estimation:** Unlike the sequential approach employed by Espasa et al. (1996) and Proietti and Pedregal (2023), where seasonality is estimated first followed by other components, our approach estimates all factors simultaneously within a general framework.

The remainder of this paper is organized as follows. Section 2 explores the "stylized facts" often observed in high-frequency data and analyzes their specific characteristics in our context of interest. Section 3 focuses on the econometric strategy to model those stylized facts. Section 4 presents an empirical application concerning daily jobless claims in Chile. Finally, Section 5 concludes the paper.

2 Stylized Facts of High-Frequency Big Data Time Series

This section outlines the characteristic features of non-financial tall big data time series.

a) Multiple Seasonal Cycles: These time series often exhibit multiple seasonal patterns (daily, weekly, monthly, and annual), which can be distorted by interactions among themselves and by the influence of weather and calendar effects. Think on the example of the introduction, about hourly retail sales data. We might observe not just a weekly cycle (higher sales on weekends) but also daily fluctuations (higher sales during lunch hours) and potentially even variations based on the day of the week (higher sales on paydays, which may be not fixed along the week or the month). Additionally, the weekly sales cycle can fluctuate depending on the season and weather. Weekends may see higher sales during periods with comfortable temperatures, while sales may to dip during extreme hot or cold spells.

b) Weather Dependence: The data may be significantly influenced by weather conditions, including precipitation, wind speed, humidity, temperature, luminosity, air pressure, cloudiness, and daylight hours. These relationships can be complex, involving multiple regimes, varying across different weather value ranges, and exhibiting dynamic, non-linear, and multiplicative structures. Furthermore, forecast weather data may be more relevant than historical observations in some cases (Cancelo et al., 2008).

   Accurately capturing weather effects requires careful consideration of the specific sector and geographical area. For broader regions, aggregated weather data is necessary, with weights assigned based on the variable's importance in each location.

   Temperature is often a primary weather factor. Categorizing temperature into neutral, cold, and hot ranges can help capture nonlinear effects. These effects can vary by time of day, day of week, and season, leading to multiple regimes. Additionally, the impact of weather can be dynamic, with past weather conditions influencing the present relationship. Nonlinear modeling techniques can be employed to capture these complex patterns.

   Daily bus ridership, for example, typically rises on rainy days as people opt for the bus over walking to avoid getting wet. However, this trend might shift on Fridays and weekends, when rain could lead to an increase in car use instead of public transportation.

c) Calendar Effects: The annual calendar composition can substantially impact the data. Key factors include:

   - Weekday vs. Weekend: The type of day (weekday or weekend) can interact with seasons and extreme weather conditions. The public transportation demand example discussed above illustrates these interactions.
   - Long Weekends: Periods preceded or followed by non-working days can also exhibit interactive effects. For instance, Thanksgiving Thursdays in the US.
   - Vacation Periods: Major holiday seasons (Christmas, Easter, summer) can influence the data. For instance, Christmas period can have a different global effect or different particular effects depending on the day of the week on which Christmas day falls.

- Special Days: Holidays falling on weekdays, strikes, elections, and seasonal transitions can have varying impacts depending on their timing.

It must be noted that aggregating the data over time, for instance, from daily to monthly level, could conceal the effects of the mentioned calendar factors, but not cancel them. Consequently, estimating these effects using aggregated monthly data can lead to inaccurate results, as we discuss in Section 3.2. Therefore, to accurately capture and model calendar effects, it is essential to work with disaggregated data.

d) Outstanding Observations: Due to the aforementioned factors, the data often contain unusual values that are not actual outliers but can be explained by econometric models incorporating weather, calendar effects, and their interactions within different regimes and structural forms. For instance, a very hot day in a weekend in summertime can generate different effects than in a weekday or a weekend in wintertime, which may appear as outliers if not properly modelled.

   While genuine outliers should be addressed, many apparent outliers can be attributed to these underlying factors, providing opportunities for causal explanations. Although estimating interaction effects with limited data points can be challenging, modeling these effects is crucial for accurate forecasting and avoiding spurious outlier corrections.

e) Importance of aggregation level: A critical consideration in analyzing such data is the level of aggregation. While weather and calendar factors exert significant influence on overall cyclical patterns, sectoral variations, labor market dynamics, and societal habits introduce disparities across sectors and geographic regions. Consequently, analyzing data at an aggregate level can obscure these heterogeneities.

   Disaggregating the data can reveal nuanced patterns, enhance model precision, and inform targeted policy interventions. However, it also introduces challenges such as increased complexity and the potential for overfitting. The optimal level of aggregation is contingent on the specific research question, data availability, and computational constraints.

f) Lack of homogeneity: As mentioned above, daily weather patterns significantly influence our high-frequency data, creating distinct daily cycles that vary across seasons and weekdays. To accurately capture these complex dynamics when working with frequencies higher than daily (for instance, hourly, half-hourly, and quarter-hourly), we propose using seasonal daily models (see also Cancelo et al., 2008). These models involve developing multiple daily models for different moments of the cycle, incorporating weather and calendar factors to explain the data's behavior.

Given the complex nature of these data, a thorough and systematic model discovery process is essential. By identifying the underlying causal factors and their intricate relationships, it is possible to explain much of the apparent volatility in the data. The paper proposes

using the automated *Autometrics* procedure (Doornik, 2009 and Hendry and Doornik, 2018) to discover parsimonious models that capture the data's characteristics. Table 1 summarizes the all the stylized facts described above.

Table 1: Stylized Facts of Tall Big Data Time Series of High Frequency

Stylized Fact	**Description**
Multiple Seasonal Cycles	Time series exhibit daily, weekly, monthly, and annual patterns, potentially influenced by interactions and weather/calendar effects.
Weather Dependence	Data is sensitive to various weather conditions, with complex relationships involving multiple regimes, value ranges, and dynamic/non-linear structures.
Calendar Effects	Annual calendar components (weekdays, long weekends, holidays, special days) impact the data, often interacting with other factors.
Outstanding Observations	Data may contain unusual values explained by model specifications including weather, calendar effects, and their interactions, rather than being true outliers.
Data Disaggregation	Analyzing data at a disaggregated level (sectoral, geographical) can reveal hidden patterns and improve model accuracy but increases complexity and computational demands.
Data Homogeneity	The relationship between data and its past/exogenous variables may not be constant over time, what may require working with a different daily model for each moment of the cycle, possibly with cross restrictions between models and with residual cross-correlations

3 Econometric strategy

Given the complex patterns influenced by calendar and weather factors in tall big data time series, analysts must consider a vast array of potential alternative exogenous variables from the outset of model building, including a possible large set of dummy variables to capture seasonal and calendar effects and their interactions. To address this challenge, we propose using an automated approach like *Autometrics* (see Doornik, 2009). This involves specifying a comprehensive initial model (the so called, general unrestricted model – GUM) formulating alternative calendar and weather-related schemes, including a possible large set of interaction terms. *Autometrics* then employs a general-to-specific strategy, conducting an exhaustive multi-path tree search –implemented by block segmentation with expansive and contracting phases– to identify the most parsimonious model while ensuring it encompasses the initial specification and meets diagnostic tests for adequacy. The algorithm also embeds Indicator Saturation Estimation (ISE), formulated in Castle et al. (2023), generalizing initial saturation indicators such as Impulse Indicator Saturation (IIS) and Step Indicator Saturation (SIS). By *saturating* the regression with impulses and step indicator variables (and other indicators

if necessary), these techniques allow the estimation of outliers (impulses) and level shifts (steps) jointly with the selection of the model (see e.g., Hendry and Doornik, 2018).

A core hyperparameter of the algorithm is the target *target size* (α), which establishes the threshold for determining the statistical significance of regressors, including impulses and step indicators. This parameter significantly influences the tree search process. Unlike backward stepwise methods, *Autometrics* explores multiple paths, employs encompassing tests against the GUM, and conducts comprehensive diagnostic checks. Consequently, while target size balances overfitting and bias, it is not the sole determinant of the final model (see Doornik, 2009).

Building upon Espasa et al. (1996), we provided a more detailed characterization of stylized facts (Table 1) commonly observed in these time series. This foundation allows analysts to construct a comprehensive set of explanatory variables that capture the complex dynamics, including nonlinearities, regime shifts, and interactions, associated with calendar and weather factors.

3.1 The algorithm

The objective is to design a methodology to estimate all seasonal, calendar, and weather effects that may be relevant to the series at hand. This goal presents two main challenges. First, we need to define a list of potentially relevant effects and their possible interactions. Second, we need to select those which are relevant, considering that the existence of actual outliers and measurement errors can distort both, the selection of the relevant effects and their quantification.

We propose the following algorithm:

1. Based on the stylized facts of Table 1, define a very general structure to capture cycles, calendar, weather effects and interactions between them. In next section we describe in detail how we define this structure for the special case of daily jobless claims in Chile. But in general it will include:

   (a) Annual, monthly, and weekly cycles. Time series with a frequency higher than daily (e.g., hourly or even higher) often exhibit daily cycles with varying patterns across different cycle phases. In these cases, employing a distinct model for each phase, such as 24 separate models for hourly data, might be necessary.

   (b) Interactions between previous cycles.

   (c) Special calendar effects such as non-labor days, Christmas, New Year, Easter, etc.

   (d) Other special effects. For example, in the empirial application in next section we include the social unrest of October 2019, the covid crisis, the regulatory change admitting claims on the internet, etc.

   (e) Weather variables, mainly to consider the effect of rain and temperature on the variable of interest.

(f) Regular and seasonal lag structure to allow for some flexibility to the deterministic seasonal effects and to model a non-seasonal cycle.

2. Starting with a big model that includes all possible effects and interactions defined above (the General Unrestricted Model—GUM), select a final model with *Autometrics* including the detection of contamination effects (outliers, location shifts, broken trends, parameter changes) using ISE. Since the number of potentially relevant regressors could be rather large, it would be important to use different target sizes for selecting outliers, deterministic effects, and the lag structure. Hence, we define an estimation procedure that consists of three steps:

   (a) Run Autometrics with ISE including all the N deterministic effects and p lags of the dependent variable in the GUM using a tight target size (α_0) to avoid keeping too many indicators. Store the selected indicators in set I_0.

   (b) Repeat step 1 with a larger, but still tight, target size (α_1) without ISE but including I_0 in the GUM. Store the selected deterministic effects in set X^*.

   (c) Run Autometrics without ISE, a standard target size (α_2), and including in the GUM the p lags, I_0, and X^*.

Defining y_t as the the logarithm of the original series, X_t as the matrix that contain all the variables finally selected after step 2.c, and $\Phi(L)$ as the corresponding autoregressive polynomial, the final model is:

$$y_t = c + \Phi(L)y_t + X_t\beta + \epsilon_t, \tag{1}$$

so that the series corrected of all the cyclical, calendar, and weather effects—which we denote as the filtered series—-would be:

$$y_t - \frac{c + X_t\beta}{1 - \Phi(L)} = \frac{\epsilon_t}{1 - \Phi(L)} \tag{2}$$

3.2 Quantitative macroeconomic leading indicators built from tall big data adjusted for calendar and weather effects

Several types of tall-big data with millions of observations per month can serve as valuable inputs for constructing quantitative macroeconomic leading indicators—typically at a monthly frequency—that offer near real-time insights. These monthly indicators can often be more reliable than qualitative indicators based on surveys of economic agents, who may lack comprehensive information or exhibit herd behavior. Examples of data suitable for constructing such indicators include electricity consumption, credit card transactions (Bodas et al., 2018), traffic data, sales, and Google's data (Choi and Varian, 2012).

It's crucial to note that economic agents' responses to calendar and weather factors may differ significantly between these high-frequency variables and broader macroeconomic

indicators like GDP. To effectively use these high-frequency variables as leading indicators, it's essential to remove the influence of calendar and weather effects.

A common–but incorrect–approach involves including monthly aggregates of weather and calendar indicators as regressors in a monthly model for the potential leading indicator (say, y_t^m) and then removing the estimated effects from y_t^m. This method fails to account for nonlinearities between these factors and the underlying daily data –say, y_t^d– used to construct y_t^m.

Consider, for example, electricity consumption as a leading economic indicator. As highlighted in previous section, its relationship with weather, particularly temperature, is nonlinear: consumption rises both with high and low temperatures but decreases in moderate conditions. For instance, a month with fluctuating temperatures—alternating hot and cold days—might exhibit high electricity consumption due to weather rather than higher cyclical demand. However, the month's average temperature could fall within a moderate range, masking the weather's impact when aggregating electricity consumption at the monthly level. This would erroneously suggest higher underlying economic activity.

The correct procedure entails estimating calendar and weather effects directly from the daily data (y_t^d) and removing these estimated effects to obtain an adjusted y_t^d series. Subsequently, aggregating this adjusted daily series yields a monthly hard indicator.

4 Empirical application to Daily Jobless Claims in Chile

4.1 Description of the data

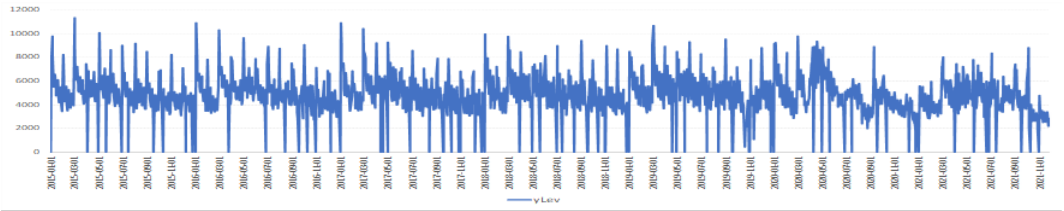
We analyze daily jobless claims in Chile collected by the Chilean Unemployment Fund Management Company (AFC, standing for *Sociedad Administradora de Fondos de Cesantía de Chile*) from January 1, 2015 to November 26, 2021 (1802 observations, excluding weekends), which exhibit significant seasonal and calendar patterns (Figures 1 and 2). Prior to April 1, 2020, claims were submitted in person, but online filing was introduced due to the COVID-19 pandemic, which began in Chile in late March 2020[1].

The agency responsible for managing these claims utilizes the data for short-term operational planning and as an early labor market indicator. Accurately quantifying the numerous seasonal and calendar effects is essential for operational planning, as they account for most of the series' short-term variability. These effects also obscure the underlying trend, hindering the use of the series as an early warning indicator.

Visual inspection and simple averages reveal pronounced seasonal and calendar patterns in the jobless claims data (Figures 3 and 4). A typical month exhibits a 50% decline in claims from the first to the last working day (Figure 4a), while a typical week shows a 30% decrease from Monday to Friday (Figure 4b). Annually, claims fall by approximately 25% from March to December. These patterns suggest a strong beginning-of-month and end-of-month effect

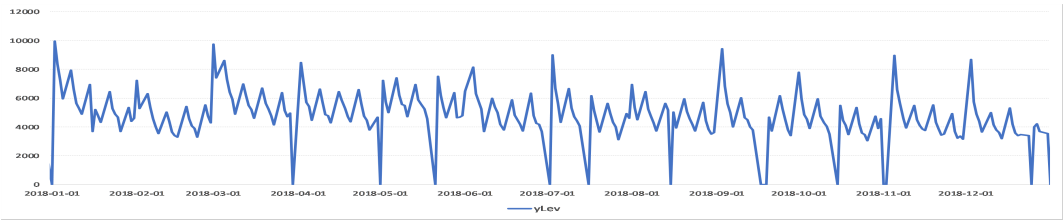
[1]These data is publicly available only at a monthly frequency. The Central Bank of Chile has access to aggregated daily information, by virtue of collaboration agreements signed with the AFC.

Figure 1: Original series (number of claims)



Source: AFC.

Figure 2: Original series (number of claims)–year 2018

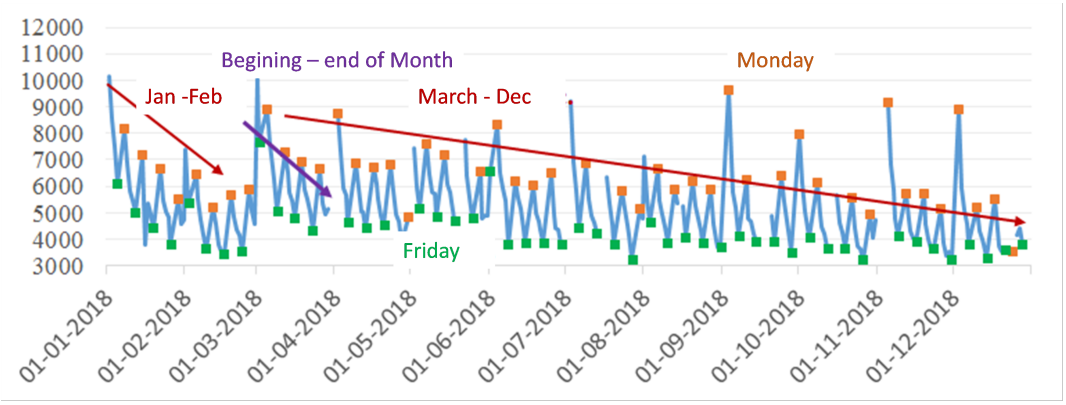


Source: AFC.

within the monthly cycle.

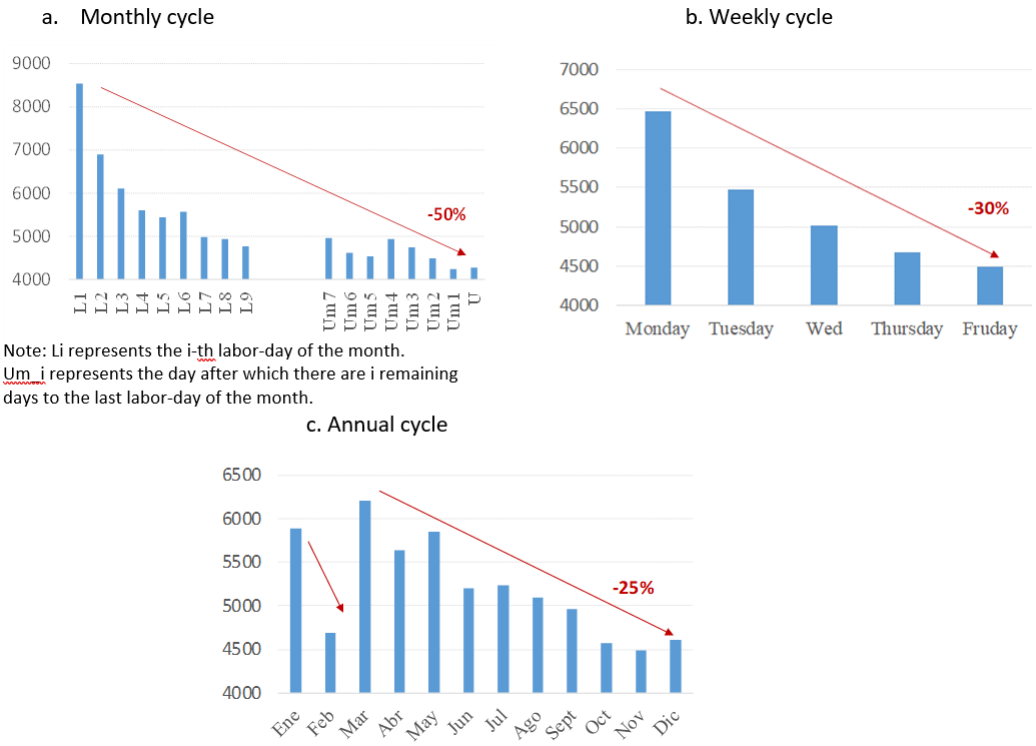
While these are the most apparent seasonal patterns, additional factors may influence jobless claims. For example, it is important to consider whether the first working day of a month falls on a Friday or Monday, account for non-working days, holidays like Christmas and New Year, and explore interactions among these factors. Furthermore, the October 2019 social unrest and the COVID-19 crisis, along with the subsequent shift to online claim filing, may have significant impacts. Weather variables, with both linear and nonlinear effects, could also interact with other factors. To comprehensively model the data, we must incorporate all these potential regressors.

Figure 3: Original series (number of claims), leaving out days with no claims- year 2018



Source: AFC.

Figure 4: Monthly, weekly, and annual averages of the original series



4.2 Implementation details

We begin with a comprehensive set of 461 potential regressors (detailed in Table 2) and incorporate a lag structure of 15 regular and seasonal lags of the dependent variable (specifically, lags 1 to 6, 9 to 11, 14 to 16, 19 to 21, 24 to 26, and 29 to 31). We employ the algorithm outlined in previous section, using target sizes α_0, α_1, and α_2 of 0.1%, 0.5%, and 1% respectively.

For including weather effects, we first construct a country-average temperature (rain) series as the average of the daily average temperature (rain) of 20 weather stations located along the country, weighted by the regional GDP corresponding to the area of influence of each station. We use a specific type of non-linear effect of temperature by partitioning the original country average into three series: one that takes the country average value when it is lower than or equal to 20 degrees and zero otherwise, another one that takes the country average value when it is higher than 20 and lower than or equal to 24 degrees and zero otherwise, and a third one for the case when the average temperature is above 24 degrees. This strategy would allow different effects of cold, mid, and hot temperatures on the number of claims. For the rain, we just include the country average.

To account for the COVID-19 pandemic, we employ the Oxford Stringency Index, measuring the intensity of mobility restrictions imposed by local authorities. This index is divided into three phases: growth (March 14, 2020 to August 6, 2020), stability (August 7, 2020 to July 17, 2021), and decline (July 19, 2021 to November 28, 2021).

The onset of the pandemic coincided with a regulatory change enabling online claim filing.

This change, combined with altered pandemic-related behavior, appears to have induced a structural shift in the seasonal pattern. To address this, we introduce a broken seasonal pattern for the COVID period by multiplying variables in lines 1 to 5 of Table 2 by an indicator equaling 1 during the COVID period and 0 otherwise. Additionally, the availability of online filing may have influenced the relevance of weather variables. Consequently, we multiply variables in line 15 of Table 2 by the COVID period indicator as well. This expansion results in a total of 840 potential regressors in the GUM, compared to the initial 461.

Table 2: Cyclical, calendar, and weather regressors included in the GUM

Deterministic and weather effects	# of Variables in the GUM active in the whole sample	# of Variables in the GUM active in the covid period	Total num of variables in the GUM	# of Variables in the final model*
Annual cycle	11	11	22	6
Monthly cycle, which differences the first and last 10 labor days of the month	20	20	40	15 (3)
Weekly cycle	4	4	8	4 (2)
Interactions between the annual cycle and the monthly cycle	240	240	480	34 (24)
Interactions between the weekly cycle and the monthly cycle	100	100	200	8 (3)
Non-labor days with 2 lags and 1 lead	4		4	2
Interactions between non-labor days (including the lead and lag) and the weekly cycle	20		20	8
Interactions between non-labor days (including the lead and lag) and the annual cycle	48		48	8
The last labor-day of the year	1		1	1
Controls for the social unrest of October 2019	2		2	2
Controls for the special national holydays around September de 18th	2		2	1
Control for December the 24th	1		1	1
Control for December the 31st	1		1	1
Controls for the Covid-19 period; the oxford stringency index broken-down into three periods (growth, stability, and decline) plus an indicator variable taking the value 1 in 2020/03/23 (when restrictions started) and -1 in 2021/11/04 (when main restrictions ended)	3		3	0
Weather variables: three temperature variables and one rain variable.	4	4	8	0
Total	**461**	**379**	**840**	**93**

Numbers in parenthesis indicate the number of variables corresponding to the covid period.

4.3 Results

Table 4 presents detailed results for the final model. From the initial 840 regressors in the GUM, Autometrics selects 93, along with 63 outlier corrections. Comparing the models with and without broken seasonal patterns, two key differences emerge: first, the number of outliers during the COVID period decreases substantially (from 76 to 43) when the broken pattern are considered. Second, the distribution of outliers within the COVID period is no longer concentrated around seasonal peaks (see Figures A.1 to A.4 in the appendix).

Figures 5 and 6 display the filtered series (equation 2) and estimated seasonal and calendar effects. Comparing the filtered and original series (first three panels of Figure 5) underscores the critical role of seasonal and calendar effects in explaining the time series' variance. Notably, even after meticulously modeling these effects, several outliers persist (last panel of Figure 5), highlighting the importance of employing our algorithm with ICE.

Additionally, Table 2 and Figure 6 demonstrate the significance of not only linear seasonal and calendar effects but also their interactions. The annual cycle (variables seasM in Table 3 and the second panel of Figure 6) exhibits the expected seasonal peaks from March to May and a decline in November, as observed in Figure 4.

The monthly cycle is characterized by a pronounced beginning-of-month effect (variables $firstDayLab$ in Table 3, and the third panel of Figure 6) and an end-of-month effect

(variables $LastDayLab$ in Table 3, and the third panel of Figure 6). The positive beginning-of-month effect spans the first seven weekdays, with the most pronounced impact occurring within the first three days, leading to daily jobless claim variations of nearly 50%. The negative end-of-month effect extends over five weekdays and is approximately one-fourth the magnitude of the beginning-of-month effect. Notably, the beginning-of-month effect diminishes by roughly 30% from March 2020 onward, coinciding with the onset of the pandemic and the introduction of online claim filing.

The weekly cycle (variables $seasD$ in Table 3 and the first panel of Figure 6) exhibits pronounced peaks at the beginning of the week. However, the magnitude of these weekly effects is somewhat attenuated post-March 2020. The Monday seasonal effect, for instance, decreases from nearly 15% in the pre-COVID period to approximately 11% after March 2020, a statistically significant decline.

Beyond these linear effects, significant interaction effects exist among the seasonal cycles. Notably, the monthly cycle interacts with both the annual cycle (variables $ProdMonth\_x\_LastD\_x$ and $ProdMonth\_x\_FirstD\_x$ in Table 3, and the fifth panel of Figure 6) and the weekly cycle (variables $ProdWeek\_x\_LastD\_x$, and $ProdWeek\_x\_FirstD\_x$ in Table 3, and the sixth panel of Figure 6). These interactions, particularly those between the monthly and annual cycles, are pronounced throughout the sample period but become more influential post-March 2020. While these interactions previously induced daily variations of approximately +-9%, they now contribute to fluctuations of around +-22% on average, with specific combinations reaching +-50%.

Interactions between the monthly and weekly cycles induce average daily variations of approximately +-7% in the pre-COVID period, increasing to +-50% after March 2020.

Non-labor days induce substantial daily variations not only on the non-labor day itself but also on the preceding and following days. These effects vary across different months (variables $LagFer$, and $FerLag\_x\_Month\_x$ in Table 3 and the seventh panel of Figure 6).

Additionally, specific events like the October 2019 social unrest (variables $SemCrisis\_18Oct$ and $Lunes21\_Oct$ in Table 3), the Chilean Independence Day holiday in September (variable $Sept\_17\_L\_J$ in Table 3), and the effects of December 24th and 31st significantly impact the series.

Regarding weather variables, neither temperature nor rainfall emerged as significant regressors in the model. While this is plausible for the post-April 2020 period with online claim filing, we anticipated some influence during earlier periods. It's possible that weather effects are indeed negligible or that our modeling approach was insufficiently nuanced. For instance, extreme weather conditions or interactions with seasonal peaks might yield relevant patterns. These questions warrant further investigation.

All in all, our findings underscore the critical importance of accurately modeling complex seasonal, calendar, and deterministic effects, along with their interactions and outliers. Without such meticulous treatment, analysis would be based on the highly volatile original series (red line in the first panel of Figure 5) rather than the stabilized filtered series (blue

line), leading to misleading conclusions.

Figure 5: Original and filtered series

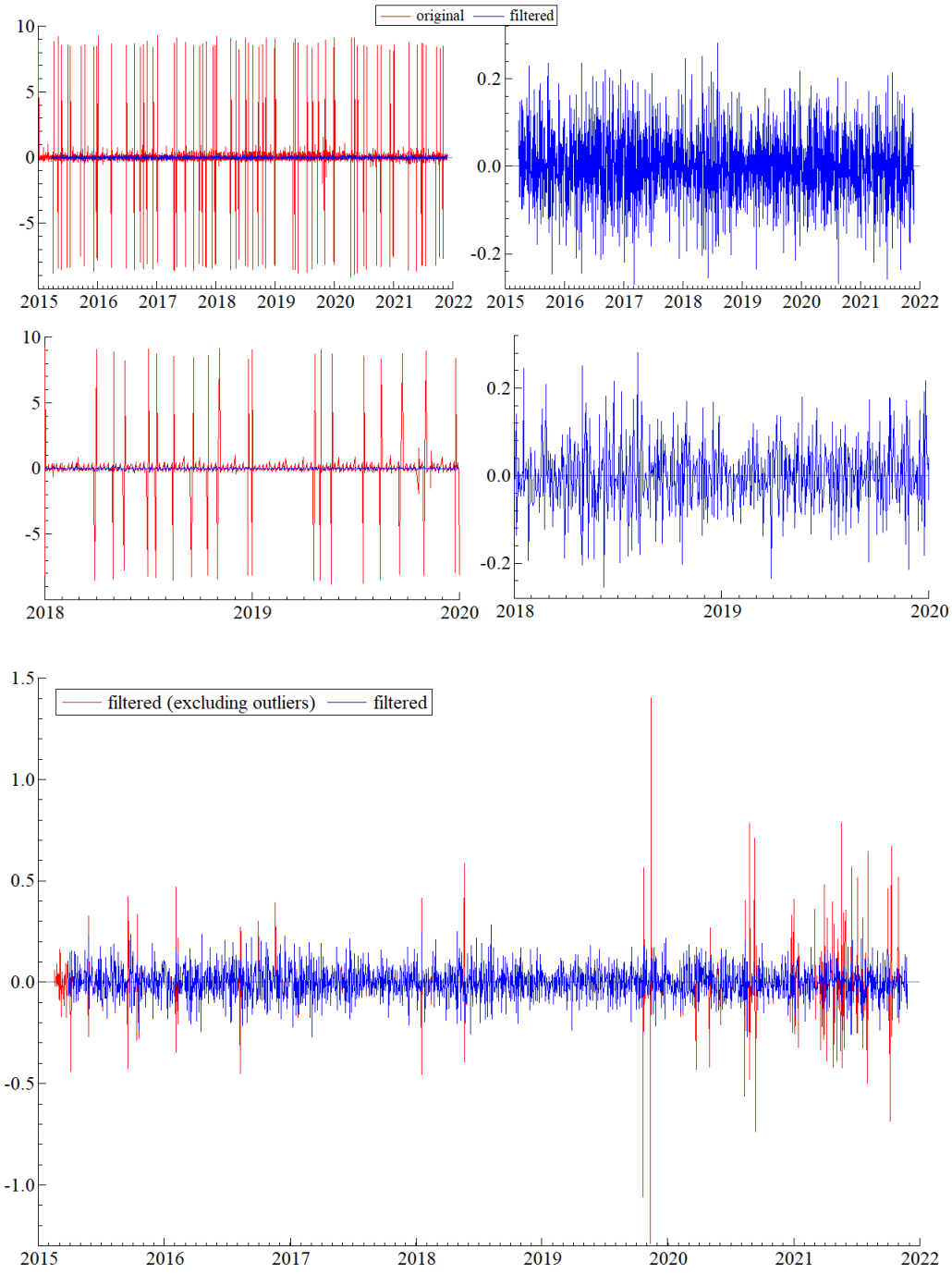


Figure 6: Seasonal effects period 2019-2021

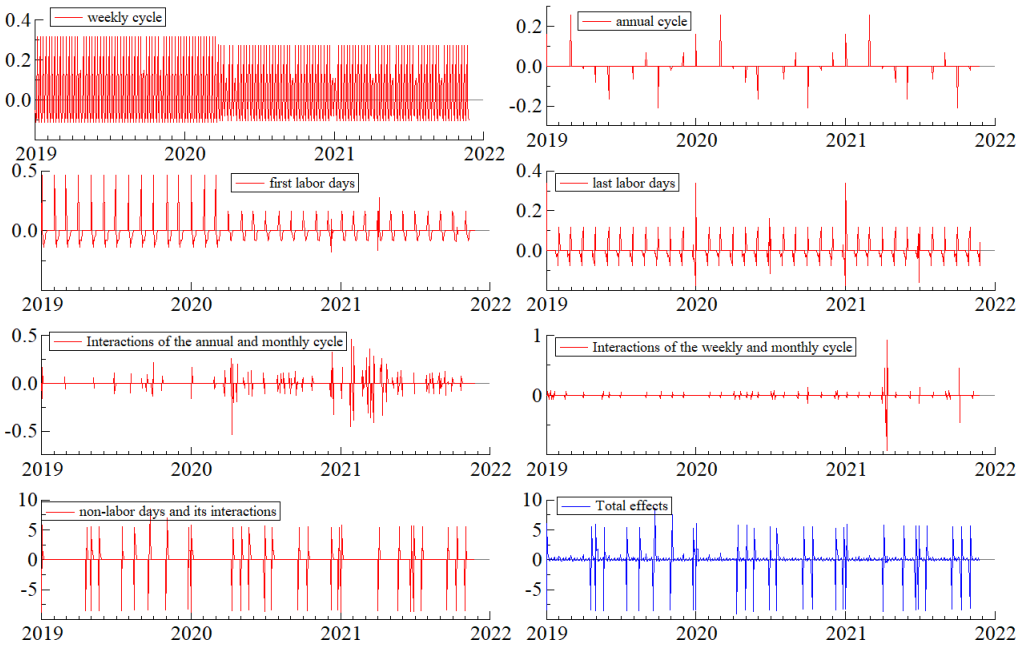


Table 3: The final model (outliers not reported)

	Coefficient	Std.Error	t-value	t-prob
dl_sol_1	-0.364074	0.01578	-23.1	0
dl_sol_2	-0.174983	0.01474	-11.9	0
dl_sol_31	-0.0019004	0.000866	-2.19	0.028
LagFer_0	-8.55081	0.01483	-577	0
LagFer_1	-3.02398	0.1335	-22.7	0
seasM_3	0.256908	0.02498	10.3	0
seasM_4	0.245613	0.02904	8.46	0
seasM_5	0.165667	0.02511	6.6	0
seasM_8	-0.0666882	0.02205	-3.02	0.003
seasM_10	-0.20885	0.03166	-6.6	0
seasM_11	-0.230909	0.02846	-8.11	0
seasD_1	0.148562	0.004455	33.3	0
seasD_2	0.0954595	0.003987	23.9	0
seasD_3	0.0199969	0.004371	4.58	0
seasD_4	-0.0944193	0.003293	-28.7	0
FerLag_0_Day_1	-0.112775	0.01794	-6.29	0
FerLag_2_Day_1	-1.4675	0.1255	-11.7	0
FerLag_-1_Day_2	-0.0599717	0.02219	-2.7	0.007
FerLag_2_Day_2	-1.45953	0.1235	-11.8	0
FerLag_2_Day_3	-1.43471	0.1266	-11.3	0
FerLag_0_Day_4	0.0971613	0.02168	4.48	0
FerLag_2_Day_4	-1.38025	0.1265	-10.9	0
FerLag_2_Day_5	-1.52108	0.1255	-12.1	0
FerLag_0_Month_1	-0.325183	0.04067	-8	0
FerLag_1_Month_4	-0.0905445	0.02735	-3.31	0.001
FerLag_1_Month_6	0.0739739	0.02704	2.74	0.006
FerLag_-1_Month_9	-0.0906634	0.02623	-3.46	6E-04
FerLag_0_Month_9	0.0870984	0.02745	3.17	0.002
FerLag_0_Month_10	0.0943003	0.02436	3.87	1E-04
FerLag_0_Month_11	0.101028	0.03493	2.89	0.004
FerLag_0_Month_12	0.150775	0.02472	6.1	0
firstDayLab1	0.465177	0.01618	28.7	0
firstDayLab2	0.477657	0.01721	27.8	0
firstDayLab3	0.395885	0.01876	21.1	0
firstDayLab4	0.258904	0.01562	16.6	0
firstDayLab5	0.178416	0.0142	12.6	0
firstDayLab6	0.0937816	0.01177	7.97	0
firstDayLab7	0.0308346	0.009153	3.37	8E-04
LastDayLab1	-0.116371	0.01381	-8.42	0
LastDayLab2	-0.160917	0.01401	-11.5	0
LastDayLab3	-0.0855029	0.01246	-6.86	0
LastDayLab4	-0.0446913	0.01109	-4.03	1E-04
LastDayLab5	-0.0300476	0.008632	-3.48	5E-04
LastDayLab_Year	-0.222226	0.04563	-4.87	0
ProdMonth_9_LastD_0	-0.21309	0.03815	-5.59	0
ProdMonth_2_LastD_1	0.0630956	0.02279	2.77	0.006
ProdMonth_6_LastD_1	0.110602	0.02324	4.76	0
ProdMonth_9_LastD_1	-0.0818583	0.02871	-2.85	0.004
ProdMonth_10_LastD_6	-0.0781186	0.02235	-3.49	5E-04
ProdMonth_9_LastD_8	-0.0873942	0.02918	-3	0.003
ProdMonth_1_FirstD_0	-0.164161	0.0323	-5.08	0
ProdMonth_9_FirstD_1	0.0497501	0.02259	2.2	0.028
ProdMonth_8_FirstD_3	-0.101347	0.02253	-4.5	0
ProdMonth_5_FirstD_4	0.054519	0.02252	2.42	0.016
ProdWeek_5_LastD_9	-0.0704004	0.01626	-4.33	0
ProdWeek_1_FirstD_0	-0.0470708	0.01413	-3.33	9E-04
ProdWeek_5_FirstD_2	0.0684154	0.01925	3.55	4E-04
ProdWeek_1_FirstD_4	-0.059566	0.01859	-3.2	0.001
ProdWeek_5_FirstD_7	0.0758822	0.01824	4.16	0
Lunes21_Oct	-2.03281	0.1139	-17.8	0
SemCrisis_18Oct	-0.295831	0.08065	-3.67	3E-04
Sept_17_L_J	-0.132755	0.0382	-3.48	5E-04
Dec24_Lab	-0.365105	0.03323	-11	0
Dec31_Lab	-0.186659	0.05142	-3.63	3E-04
seasDCovid_1	-0.0411202	0.00605	-6.8	0
seasDCovid_4	0.0350613	0.005959	5.88	0
firstDayLabCovid1	-0.301026	0.01954	-15.4	0
firstDayLabCovid2	-0.197139	0.02236	-8.82	0
firstDayLabCovid3	-0.116716	0.01965	-5.94	0
ProdMonth_9_LastD_Covid_0	0.157014	0.06029	2.6	0.009
ProdMonth_5_LastD_Covid_1	0.132348	0.04689	2.82	0.005
ProdMonth_7_LastD_Covid_1	0.0860275	0.04737	1.82	0.07
ProdMonth_8_LastD_Covid_1	0.109042	0.04164	2.62	0.009
ProdMonth_2_LastD_Covid_3	0.131972	0.05701	2.31	0.021
ProdMonth_1_LastD_Covid_4	-0.456598	0.05737	-7.96	0
ProdMonth_3_LastD_Covid_6	-0.130973	0.05067	-2.58	0.01
ProdMonth_4_LastD_Covid_6	-0.197764	0.04623	-4.28	0
ProdMonth_5_LastD_Covid_6	0.119561	0.04087	2.93	0.004
ProdMonth_3_LastD_Covid_7	-0.413593	0.06255	-6.61	0
ProdMonth_9_LastD_Covid_8	0.0992232	0.0535	1.85	0.064
ProdMonth_2_FirstD_Covid_0	-0.386797	0.0591	-6.54	0
ProdMonth_3_FirstD_Covid_3	0.264845	0.06694	3.96	1E-04
ProdMonth_3_FirstD_Covid_4	0.312208	0.06643	4.7	0
ProdMonth_4_FirstD_Covid_4	0.160307	0.05404	2.97	0.003
ProdMonth_12_FirstD_Covid_4	-0.136617	0.06009	-2.27	0.023
ProdMonth_4_FirstD_Covid_5	0.28319	0.06034	4.69	0
ProdMonth_12_FirstD_Covid_5	-0.141848	0.05911	-2.4	0.017
ProdMonth_4_FirstD_Covid_6	0.539222	0.06527	8.26	0
ProdMonth_4_FirstD_Covid_7	0.202431	0.04686	4.32	0
ProdMonth_12_FirstD_Covid_7	0.327222	0.06008	5.45	0
ProdMonth_9_FirstD_Covid_8	0.127779	0.04664	2.74	0.006
ProdMonth_3_FirstD_Covid_9	0.362693	0.05701	6.36	0
ProdMonth_8_FirstD_Covid_9	0.110984	0.04063	2.73	0.006
ProdWeek_3_LastD_Covid_0	-0.137297	0.04046	-3.39	7E-04
ProdWeek_3_FirstD_Covid_3	0.452589	0.05191	8.72	0
ProdWeek_1_FirstD_Covid_6	-0.92635	0.0842	-11	0

5 Conclusions

This paper explores the characteristics of non-financial, high-frequency tall big data time series. We extend Hendry (2015b) concept of "tall" big data to encompass time series with numerous observations, few endogenous variables, and potentially hundreds of exogenous ones requiring careful formulation and identification.

These high-frequency time series often exhibit complex and changing seasonal and high-frequency patterns, which can lead to misleading outlier identification. To address this, we propose a set of stylized facts to characterize these data and guide model formulation, considering regime-switching, nonlinearities, and multiplicative effects. The search for optimal model structures can be efficiently conducted using *Autometrics*.

Another important consideration lies in the need for weather variables and calendar-related dummy variables when modeling high-frequency data. These factors, including weekdays, weekends, holidays, special days, and tehir interactions between eachother and with weather variables can significantly influence the data. Accurate modeling of these effects is crucial for policy-making and can help differentiate genuine outliers from data anomalies.

To build appropriate models, we advocate for a two-step approach: defining relevant exogenous variables and employing *Autometrics* for model selection. Compared to alternative methods, our approach offers the advantages of enabling causal interpretations of influential factors and simultaneously estimating all seasonal components, rather than relying on sequential procedures.

Our application to daily jobless claims in Chile demonstrates the importance of a comprehensive initial model (GUM) that includes numerous dummy variables to capture complex effects of seasonality, calendar and other deterministic effects, the interactions between them, and outlying observations. The initial model includes 841 regressors and after a general-to-specific modelling process the final model keeps 93 regressors not counting outliers (plus 63 outliers), many of them referring to interactive effects. These results emphasize the importance of capturing interactions between different seasonal cycles and calendar factors. Neglecting these interactions can lead to misleading conclusions.

References

Bodas, D., J. García, J. Murillo, M. Pacce, T. Rodrigo, P. Ruiz de Aguirre, C. Ulloa, J. d. D. Romero, and H. Valero (2018). Measuring retail trade using card transactional data. Technical Report 18/3, BBVA Research, Madrid.

Cancelo, J., A. Espasa, and R. Grafe (2008). Forecasting the electricity load from one day to one week ahead for the spanish system operator. *International Journal of Forecasting 24*(4), 588–602.

Castle, J. L., J. A. Doornik, and D. F. Hendry (2023). Robust discovery of regression models. *Econometrics and Statistics 26*, 31–51.

Cavallo, A. and R. Rigobon (2016). The billion prices project: Using online prices for measurement and research. *Journal of Economic Perspectives 30*(2), 151–178.

Choi, H. and H. Varian (2012). Predicting the present with google trends. *Economic record 88*, 2–9.

Den Boer, A. V. (2015). Dynamic pricing and learning: historical origins, current research, and new directions. *Surveys in operations research and management science 20*(1), 1–18.

Doornik, J. (2009). Autometrics. In J. Castle and N. Shephard (Eds.), *The Methodology and Practice of Econometrics*. Oxford University Press.

Dutta, G. and K. Mitra (2017). A literature review on dynamic pricing of electricity. *Journal of the Operational Research Society 68*(10), 1131–1145.

Espasa, A., J. Revuelta, and J. Cancelo (1996). Automatic modelling of daily series of economic activity. In A. Prat (Ed.), *Proceedings in Computational Statistics*, pp. 51–64. Heidelberg: Physica Verlag.

Franke, B. and et al (2016). Statistical inference, learning and models in big data. *International Statistical Review*, 1–19.

Hendry, D. (2015a). *Introductory Macro-econometrics: a new approach.* Timberlake Consultants Ltd.

Hendry, D. (2015b). Mining big data by statistical methods. *The European Financial Review*, 69–72.

Hendry, D. and J. Doornik (2018). *Empirical Econometric Modelling using PcGive* (8th ed.). Timberlake Consultants Press.

Proietti, T. and D. J. Pedregal (2023). Seasonality in high frequency time series. *Econometrics and Statistics 27*, 62–82.

Saharan, S., S. Bawa, and N. Kumar (2020). Dynamic pricing techniques for intelligent transportation system in smart cities: A systematic review. *Computer Communications 150*, 603–625.

Zhou, M., D. Wang, Q. Li, Y. Yue, W. Tu, and R. Cao (2017). Impacts of weather on public transport ridership: results from mining data from different sources. *Transportation Research Part C: Emerging Technologies 75*(3), 17–29.

Appendix A Appendix

Figure A.1: Distribution of outliers along the year. Comparison of the Covid period vs. the rest of the sample

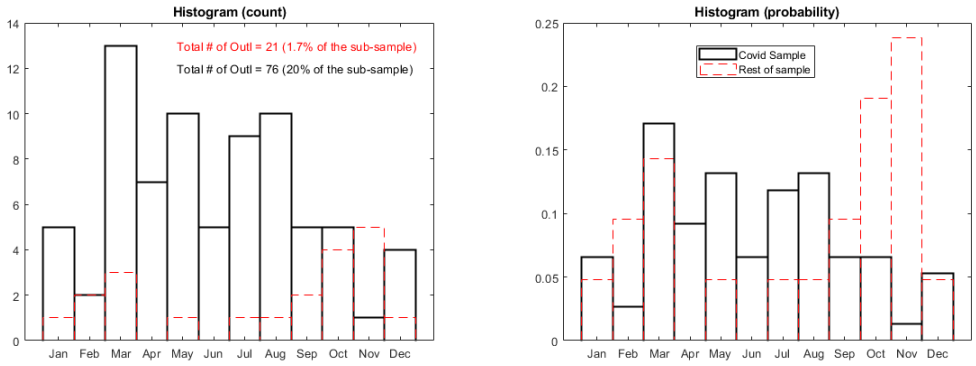


Figure A.2: Distribution of outliers along the month. Comparison of the Covid period vs. the rest of the sample

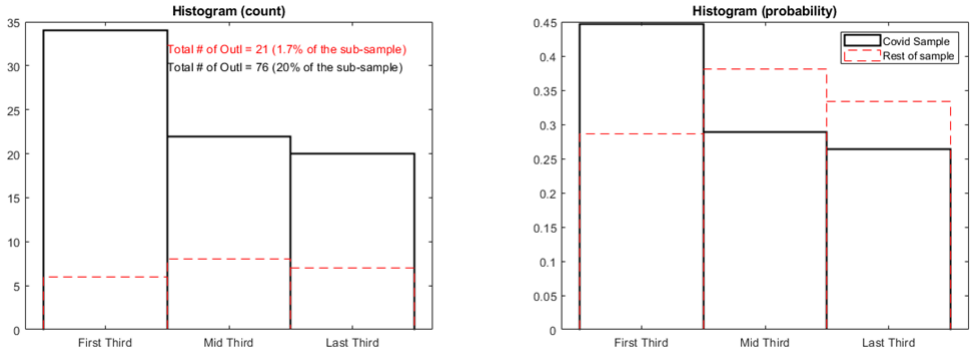


Figure A.3: Distribution of outliers along the week. Comparison of the Covid period vs. the rest of the sample

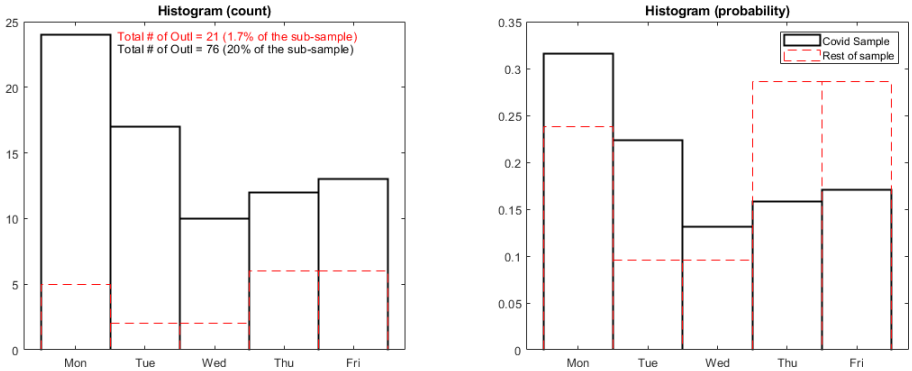
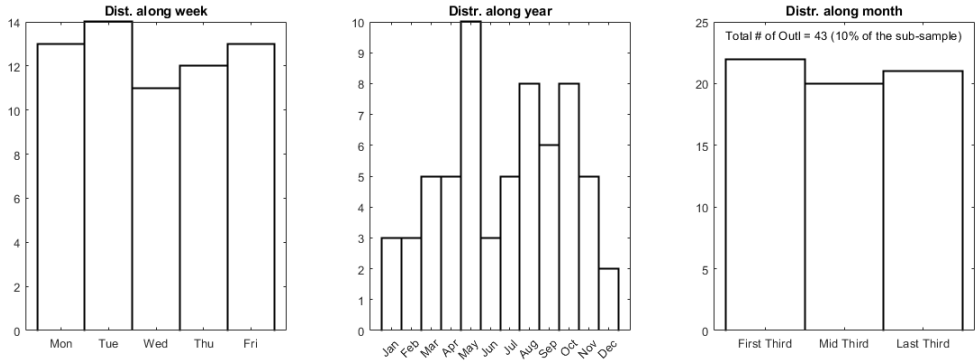


Figure A.4: Distribution of outliers in the Covid period after allowing broken seasonal effects.



Documentos de Trabajo Banco Central de Chile	Working Papers Central Bank of Chile
NÚMEROS ANTERIORES	PAST ISSUES
La serie de Documentos de Trabajo en versión PDF puede obtenerse gratis en la dirección electrónica: www.bcentral.cl/esp/estpub/estudios/dtbc. Existe la posibilidad de solicitar una copia impresa con un costo de Ch$500 si es dentro de Chile y US$12 si es fuera de Chile. Las solicitudes se pueden hacer por fax: +56 2 26702231 o a través del correo electrónico: bcch@bcentral.cl.	Working Papers in PDF format can be downloaded free of charge from: www.bcentral.cl/eng/stdpub/studies/workingpaper. Printed versions can be ordered individually for US$12 per copy (for order inside Chile the charge is Ch$500.) Orders can be placed by fax: +56 2 26702231 or by email: bcch@bcentral.cl.

DTBC – 1023
Modelling high frequency non-financial big time series with an application to jobless claims in Chile.
Antoni Espasa, Guillermo Carlomagno

DTBC – 1022
Aggregating Distortions in Networks with Multi-Product Firms
Yasutaka Koike-Mori, Antonio Martner

DTBC – 1021
Análisis de redes aplicado al sistema de pagos de alto valor del BCCh
Álvaro González, Carmen López, María José Meléndez

DTBC – 1020
Financial advisory firms, asset reallocation and price pressure in the FOREX market
Francisco Pinto-Avalos, Michael Bowe, Stuart Hyde

DTBC – 940* (Revised)
Overborrowing and Systemic Externalities in the Business cycle Under Imperfect Information
Juan Herreño, Carlos Rondón-Moreno

DTBC – 1019
Through Drought and Flood: the past, present and future of Climate Migration
Elías Albagli, Pablo García Silva, Gonzalo García-Trujillo, María Antonia Yung

DTBC – 1018
Supply Chain Uncertainty and Diversification
Ignacia Cuevas, Thomas Bourany, Gustavo González

DTBC – 1017
Is the Information Channel of Monetary Policy Alive in Emerging Markets?
Mariana García-Schmidt

DTBC – 1016
The Portfolio Choice Channel of Wealth Inequality
Mauricio Calani, Lucas Rosso

DTBC – 1015
Fiscal Consolidations in Commodity-Exporting Countries: A DSGE Perspective
Manuel González-Astudillo, Juan Guerra-Salas, Avi Lipton

DTBC – 1014
Accounting for Nature in Economic Models
Nicoletta Batini, Luigi Durand

DTBC – 1013
Transmission Mechanisms in HANK: an Application to Chile
Benjamín García, Mario Giarda, Carlos Lizama, Ignacio Rojas

DTBC – 1012
Cyclical wage premia in the informal labour market: Persistent and downwardly rigid
Daniel Guzmán

DTBC – 1011
Macro Implications of Inequality-driven Political Polarization
Alvaro Aguirre

DTBC – 1010
Firm Shocks, Workers Earnings and the Extensive Margin
Álvaro Castillo, Ana Sofía León, Antonio Martner, Matías Tapia

DOCUMENTOS DE TRABAJO Octubre 2024